ISSN: 0974 – 3987
IJBST (2011), 4(4):23-28

Functional genomics and reverse vaccinology approach to identify better vaccine for tuberculosis

P. Vijayalakshmi[1], G. Ramesh Kumar[2], C.P. Rajadurai[2], P. Daisy[1*]

1 Department of Bioinformatics Holy cross college, Trichy, INDIA
daisylesslie@gmail.com
2 AU-KBC Research Centre MIT Campus, Anna University, Chromepet, INDIA

ABSTRACT:
Tuberculosis is an infectious disease that has plagued humans which is caused by *Mycobacteriam tuberculosis*. The study of functional genomics on *M.tuberculosis* CDC1551 is to find the non-coding functional elements of the genome regions simply called as hypothetical protein which function has to be predicted with the higher level of accuracy or confidence level by using different comparative and functional genomics tools. By comparing the genome sequence of *M. tuberculosis* with vaccine regions present in other micro-organism, the sequences can be correlated so as to gain the gene patterns may act as vaccine. This was manually reannotated using functional genomics tools to identify the functions for missing ORF's in the genome of *M. tuberculosis.* Further, it was traced out for the epitope or antigenic regions like TAP, HLA, Proteosomal cleavage site and MHC using vaccination search tools. NP_337985.1, a hypothetical protein sequence function was predicted as FAD/FMN-containing dehydrogenase having the entire epitope region with high scoring value

Key words: *Mycobacteriam tuberculosis*, Functional genomics, Protein modeling and Reverse vaccinology

INTRODUCTION

Tuberculosis (TB) describes an infectious disease that has plagued humans since the Neolithic times which is caused by *Mycobacterium tuberculosis*. Streptomycin, the first antibiotic to fight TB, was introduced in 1946, and isoniazid (Laniazid, Nydrazid) became available in 1952. *M.tuberculosis*, along with *M. bovis*, *M. africanum*, and *M. microti* all cause the disease known as tuberculosis (TB) and are members of the tuberculosis species complex. Each member of the TB complex is pathogenic, but *M. tuberculosis* is pathogenic for humans [1].

Genome information of Mycobacterium tuberculosis

The original sequence and annotation of Mycobacterium tuberculosis strain CDC1551 identified 4293 genes (Cole et al., 1998). This included 4184 genes thought to encode proteins and 48encoding stable RNA and 56 encode the pseudogenes. GC content of this strain is 65% and percentage of coding region is 90%.The current nucleotide sequence now contains 4,403,837 nt**.**

Functional Genomics

Functional genomics studies the function (coded proteins), expressions and regulation from genes as the interaction better than different genes; it requires analysing the total protein produced by the genes. Functional Genomics is therefore not simply a process towards novel drug discovery, but a general approach to assign biological functions to genes with currently unknown roles in all organisms**.**

Vaccine

Vaccine is an immuno-biological substance designed to produce specific protection against a disease. Immunization not only protects the individual against infection but, if high levels of vaccination are maintained, can prevent or contain epidemics, or even eradicate diseases entirely [2].

Reverse vaccinology is an improvement on vaccinology, pioneered by Rino Rappuoli and first used against meningococcus. Since then, it has been used on several other organisms. Reverse vaccinology is built on genome-based antigen discovery and has largely replaced classical vaccinology methods based on growing and dissecting the microorganism. The main advantage of the approach is the fast prediction of vaccine candidates.

In our present study, investigation has been done to find the suitable new vaccine for tuberculosis disease through Reverse vaccinology approach. We used several vaccine region prediction tools that are user-friendly.

MATERIALS AND METHODS

Complete hypothetical genome sequence of mycobacterium tuberculosis strain CDC1551 were obtained from the NCBI and its function was annotated by using BLAST, COG, BLOCKS ,SCANPROSITE, PRODOM.

VACCINE TOOLS

NetCTL 1.0 server

NetCTL 1.0 server predicts CTL epitopes in protein sequences. The server allows for predictions of CTL epitopes restricted to 10 MHC supertype.

ANTIGENIC EMBOSS

Antigenic predicts potentially antigenic regions of a protein sequence, using the method of Kolaskar and

Tongaonkar. Application of this method to a large number of proteins has shown that their method can predict antigenic determinants with about 75% accuracy which is better than most of the known methods. This method is based on a single parameter and thus very simple to use.

PAPROC (Prediction Algorithm for proteasomal Cleavage)
PAProC (Prediction Algorithm for Proteasomal Cleavages), a public prediction tool for proteasomal cleavages. PAProC offers information on both the general cleavability of amino acid sequences (cuts per amino acids) and individual cleavages (positions and estimated strength).

TAP PRED
TAPPred is an on-line service for predicting binding affinity of peptides toward the TAP transporter. The prediction of TAP binding peptides is crucial in identifying the MHC class-1 restricted T cell epitopes. The Prediction is based on cascade SVM, using sequence and properties of the amino acids[3].

MHC II Binding prediction
The identification of MHC class II restricted peptide epitopes is an important goal in immunological research.

MHC I Binding prediction
Several accurate prediction systems have been developed for prediction of class I major histocompatibility complex (MHC). The predictions are based on artificial neural networks trained on data from 55 MHC alleles (43 Human and 12 non-human), and position-specific scoring matrices (PSSMs) for additional 67 HLA alleles [4].

MHC I Processing Prediction
Epitopes presented by major histocompatibility complex (MHC) class I molecules are selected by a multi-step process. The first computational prediction of this process based on in vitro experiments characterizing proteasomal cleavage, transport by the transporter associated with antigen processing (TAP) and MHC class I binding[5].

FDR4 (Affinity for HLA)
This server FDR4 is meant for the prediction of binding affinity of peptide binders in an antigenic sequence for a MHC class II allele HLA-DRB1*0401. Methods developed in the past can only predict whether a peptide is a binder or non-binder of this allele.

HLA –AFFINITY
The preliminary requirement for the stimulation of cytotoxic T cell response, a mechanism against viruses and certain tumors, is the processing and presentation of endogenous antigenic peptides by MHC-I molecules on the surface of the cell. Methods have been developed to classify and predict the binders and non-binders of MHC[6].

Pro pred MHC Class-II Binding Peptide Prediction
The aim of this server is to predict MHC Class-II binding regions in an antigen sequence, using quantitative matrices. MHC molecules are cell surface glycoproteins, which take active part in host immune reactions[7].

HLA_BIND: Prediction of MHC type I (HLA) peptide binding
This Web site allows users to locate and rank 8-mer, 9-mer, or 10-mer peptides that contain peptide-binding motifs for HLA class I molecules.

Modeller 9v2
MODELLER 9v2 is used for homology or comparative modeling of protein three-dimensional structures. The user provides an alignment of a sequence to be modeled with known related structures and MODELLER automatically calculates a model containing all non-hydrogen atoms[8].

SWISS-PDB Viewer
Swiss-PdbViewer is tightly linked to SWISS-MODEL, an automated homology modeling server. Deep View allows to build models from scratch, simply by giving an amino-acid sequence. Deep View can find hydrogen bonds within proteins and between proteins and ligandsSwiss-PdbViewer.

PyMOL
PyMOL is a molecular viewer developed in the spirit of RasMol and Open RasMol and intended for visualization of 3D chemical structures including X-ray crystal structures of: proteins, nucleic acids (DNA, RNA, & tRNA), and carbohydrates, as well as small molecule structures of drug leads, inhibitors, metabolites, sugars, nucleoside phosphates, and other ligands including inorganic salts and solvent molecules.

RESULTS

Functional Genomics results

M.tuberculosis was taken as a pathogenic bacteria and functional annotation was performed. 2042 gene sequence were found to be present with unknown functions. Out of 2042 unknown sequences 1070 sequences was taken for functional annotation. Table 1 shows the functional annotation results using functional genomics tools.

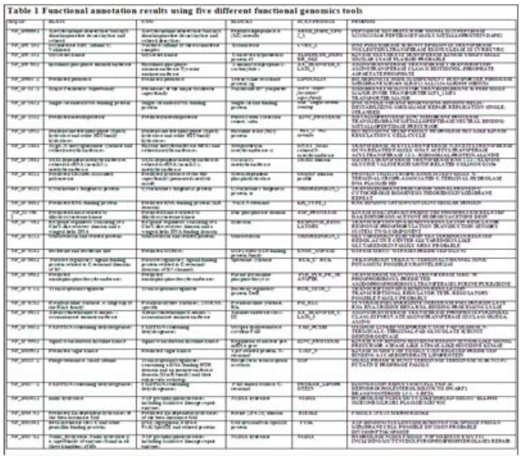
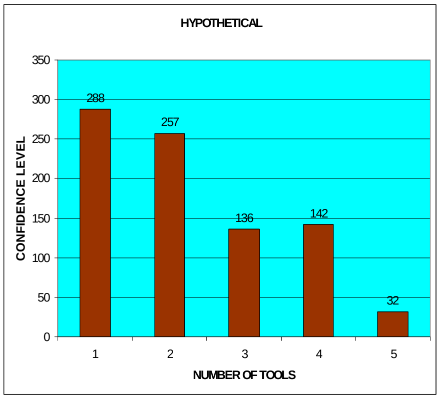
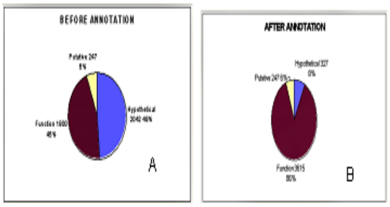


Table 1 Functional annotation results using five different functional genomics tools



Graph1: Confidence level for 1070 unknown sequence of Mycobacterium tuberculosis

After functional annotation, functions for 1070 hypothetical sequences (Graph 1) had been found. Further, the functional similarity had been checked for those hypothetical sequences based on the confidence level (i.e.) Percentage of occurrence of positive results / Number of tools used. Graph- 2 shows the percentage of hypothetical, putative and functional sequence in *M. tuberculosis* during before annotation and after annotation.



Graph 2 : Percentage of functional and non-functional sequences (a) Before annotation (b) after annotation.

VACCINE TOOL RESULT

32 protein sequences with similar function were taken for the vaccine studies. Each sequences was submitted to different antigenic tools including, antigenic EMBOSS, TAP PRED,NetCTL-1.0 Server Prediction, MHC-1 Binding Predictions, MHC-II binding predictions, MHC-I processing predictions, FDR4, PROPRED, MHC Class-II Binding Peptide Prediction HLA-BIND, PAPROC and HLA-A2. The score value for each submitted sequence were observed and the sequences with high score value were shortlisted and presented in table 2-11 which contains results from all 10 vaccine Tools of submitted sequences.

Table2: Net CTL Prediction Result

Sequence I.D	MHC binding affinity	TAP transport efficiency	Prediction score
gi_15841632	0.2678	-2.3980	1.7264
gi_15841682	0.1480	3.2530	1.2674
gi_15841696	0.2169	0.9140	1.6189
gi_15841723	0.1477	-0.3950	0.9907
gi_15841900	0.2470	2.6430	1.8962
gi_15841980	0.4756	0.1930	3.3381
gi|15842004	No result	No result	No result
gi_15842113	0.5198	3.0500	3.6903
gi_15842153	0.3332	0.1340	2.3685
gi_15842207	0.4753	3.1060	3.4833
gi_15842227	0.4074	2.8810	3.0067
gi_15842364	0.3564	3.0750	2.6737
gi_15842416	0.3446	0.9120	2.4847
gi|15842451	No result	No result	No result
gi_15842528	0.1894	0.6570	1.4169
gi_15842719	0.4377	3.0270	3.2191
gi_15842784	0.1179	3.0620	1.0531
gi|15842817	No result	No result	No result
gi_15842826	0.1637	0.2530	1.2079
gi_15842831	0.0787	2.5960	0.7640
gi_15842838	0.1576	0.8900	1.2135
gi|15842891	No result	No result	No result
gi_15842924	0.2894	2.5090	2.1758
gi_15842948	0.1446	-0.0050	1.0753
gi_15842951	0.2418	3.0530	1.8877
gi_15842826	0.1637	0.2530	1.2079
gi_15843354	0.3511	2.2210	2.5955
gi_15843442	0.5749	2.7560	**4.1419**
gi_15843502	0.4747	2.5900	3.4347

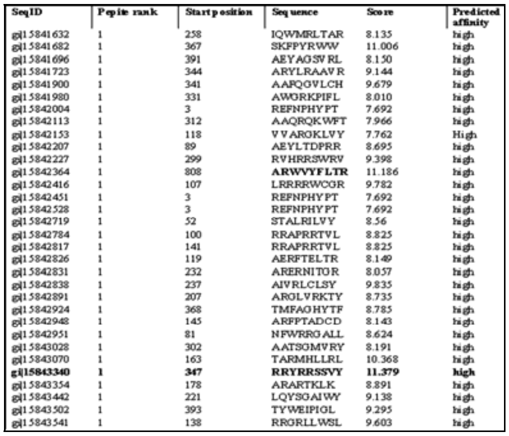
Table 3: Antigenic EMBOSS Result

Sequence I.D	score	length	residue	sequence
NP_336669.1	1.221	21	311->331	KANVVPATAEAVVDCRVLPGR
NP_336719.1	1.186	13	193->205	RATVDVLHALIER
NP_336733.1	1.207	37	299->335	LHGKVVGAIAAAARPLAIPVIVLAGQVSLDKSALRSA
NP_336760.1	1.207	10	170->179	LPDVLVLRSL
NP_336937.1	**1.269**	**15**	**291->305**	**FFVAAFQGVLCLFLL**
NP_337017.1	1.26	22	260->281	MSSCIVAAQVVMVPVAYVVGTR
NP_337041.1	1.238	15	69->83	GKGAPVIVVGHVYTS
NP_337150.1	1.299	24	32->55	GGLLVVVVAMLLGVDPGGVLSQQP
NP_337190.1	1.245	12	148->159	LAEHLHVHVVPR
NP_337244.1	1.186	34	6->39	ELAAVAARTFPLACPPAVAPEHIASFVDANLSSA
NP_337264.1	1.228	34	398->431	DIGLYRGHGYAVEKIKVFDAFPLTHYVECVALLT
NP_337401.1	1.231	11	439->449	ECSVCHTVNRT
NP_337453.1	1.245	26	252->277	LGLALGVLYVPCAGPILAAIVVAGAT
NP_337488.1	1.218	13	4->16	VVVDAVEHLVRGI
NP_337565.1	1.201	9	4->12	RRCVVVGTA
NP_337756.1	1.225	15	80->94	KDELASCPPILVLTG
NP_337821.1	1.176	13	11->23	TSWCGYCLRLKTA
NP_337854.1	1.247	20	113->132	ERVVVANADQLLIVVALADP
NP_337863.1	1.173	16	114->129	ASIVAIVRDEDVLASP
NP_337868.1	1.256	19	26->44	TWCVLDLVLPLECGGCGAP
NP_337875.1	1.211	18	185->202	NVLSRAIVRLCLSYVSMP
NP 337928.1	1.178	28	6->33	EDRLLSVHDVLGPVRVRLLGGSVLA
NP_337961.1	1.197	16	59->74	YLDALSGLFVVQVGYG
NP_337985.1	1.175	19	102->120	TADCDVVRVDFAPSAAAQV
NP_337988.1	1.208	21	39->59	GALLIGIGVGVAAVLRLVLSE
NP_338065.1	1.223	30	180->209	LKPVHALADCGRVVLVDIGLDLAHTDVLGF
NP_338107.1	1.208	21	70->90	RALLSAYCETWSVYVAAVQRV
NP_338377.1	1.215	43	4->46	TRVYAVPVPGSAQSAYACGVERLLASYRSIPATASIRLAKPTS
NP_338391.1	1.226	23	4->26	LSAGVLLYRARAGVVDVLLAHFG
NP_338479.1	1.189	16	48->63	LGECDYLYVSHLHKDH
NP_338539.1	1.229	17	468->484	GFDVVLLVDDWHMIVGA

Table 4: PaProC Tool Result

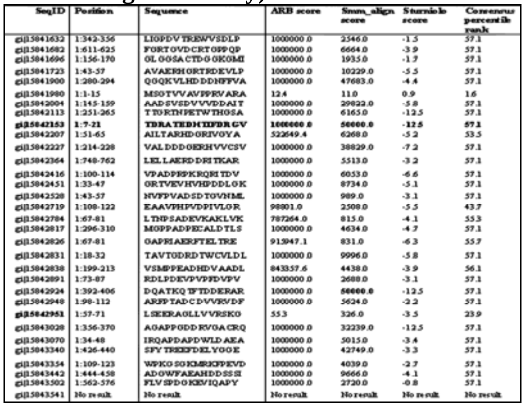
Seq ID	Position	Aminoacid	Cleavage Strength	Cleavage prediction
gi|15841632	189	L	257.402794665908	+++
gi|15841682	116	F	213.8095857559	+++
gi|15841696	241	V	183.68067514736	+++
gi|15841723	217	L	249.996055017808	+++
gi|15841900	398	L	250.92154763626	+++
gi|15841980	395	Y	269.315923315108	+++
gi|15842004	85	Y	215.65339683	+++
gi|15842113	139	F	178.1607714081	+++
gi|15842153	84	L	293.1820017735	+++
gi|15842207	80	L	208.2192515732	+++
gi|15842227	59	G	208.8292765907	+++
gi|15842364	744	I	236.950150778208	+++
gi|15842416	228	E	262.9769537558	+++
gi|15842451	7	D	182.4383937985	+++
gi|15842528	29	V	162.8022188001	+++
gi|15842719	92	L	349.6701934764	+++
gi|15842784	67	L	210.2590879412	+++
gi|15842817	157	L	305.350834874	+++
gi|15842826	20	Y	221.2539072725	+++
gi|15842831	67	S	147.257148666	+++
gi|15842838	92	G	222.339137943508	+++
gi|15842891	161	R	202.1893057558	+++
gi|15842924	293	A	275.8992279054	+++
gi|15842948	34	L	196.145104061368	+++
gi|15842951	87	I	186.437379724	+++
gi|15843028	42	L	222.3296165297	+++
gi|15843070	61	R	189.334311651	+++
gi|15843340	393	W	276.9141733054	+++
gi|15843354	55	A	180.1016043278	+++
gi|15843442	436	L	252.1533971245	+++
gi|15843502	31	P	264.990672265	+++
gi|15843541	181	L	217.521978341	+++

Table 5: TAP PRED Result



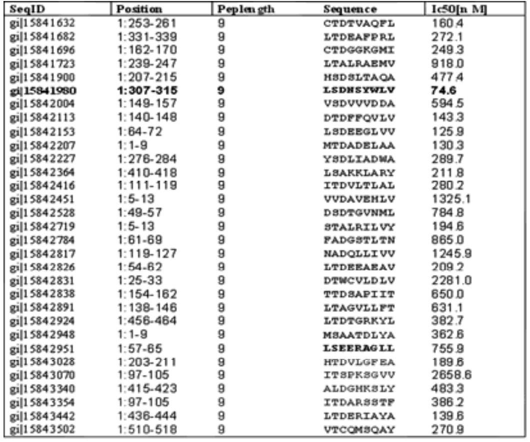
SeqID	Pepite rank	Start position	Sequence	Score	Predicted affinity
gi|15841632	1	258	IQWMRLTAR	8.135	high
gi|15841682	1	367	SKFPYRWW	11.006	high
gi|15841696	1	391	AEYAGSVRL	8.150	high
gi|15841723	1	344	ARYLRAAVR	9.144	high
gi|15841900	1	341	AAFQGVLCH	9.679	high
gi|15841980	1	331	AWGRKPIFL	8.010	high
gi|15842004	1	3	REFNPHYPT	7.692	high
gi|15842113	1	312	AAQRQKWFT	7.966	high
gi|15842153	1	118	VVARGKLVY	7.762	High
gi|15842207	1	89	AEYLTDPRR	8.695	high
gi|15842227	1	299	RVHRRSWRV	9.398	high
gi|15842364	1	808	**ARWVYFLTR**	11.186	high
gi|15842416	1	107	LRRRRWCGR	9.782	high
gi|15842451	1	3	REFNPHYPT	7.692	high
gi|15842528	1	3	REFNPHYPT	7.692	high
gi|15842719	1	52	STALRILVY	8.56	high
gi|15842784	1	100	RRAPRRTVL	8.825	high
gi|15842817	1	141	RRAPRRTVL	8.825	high
gi|15842826	1	119	AERFTELTR	8.149	high
gi|15842831	1	232	ARERNITGR	8.057	high
gi|15842838	1	237	AIVRLCLSY	9.835	high
gi|15842891	1	207	ARGLVRKTY	8.735	high
gi|15842924	1	368	TMFAGHYTF	8.785	high
gi|15842948	1	145	ARFPTADCD	8.143	high
gi|15842951	1	81	NFWRRGALL	8.624	high
gi|15843028	1	302	AATSGMVRY	8.191	high
gi|15843070	1	163	TARMHLLRL	10.368	high
gi|15843340	**1**	**347**	**RRYRRSSVY**	**11.379**	**high**
gi|15843354	1	178	ARARTKLK	8.891	high
gi|15843442	1	221	LQYSGAIWY	9.138	high
gi|15843502	1	393	TYWEIPIGL	9.295	high
gi|15843541	1	138	RRGRLLWSL	9.603	high

Table 6: MHC-II binding Prediction Result (High score have good affinity)



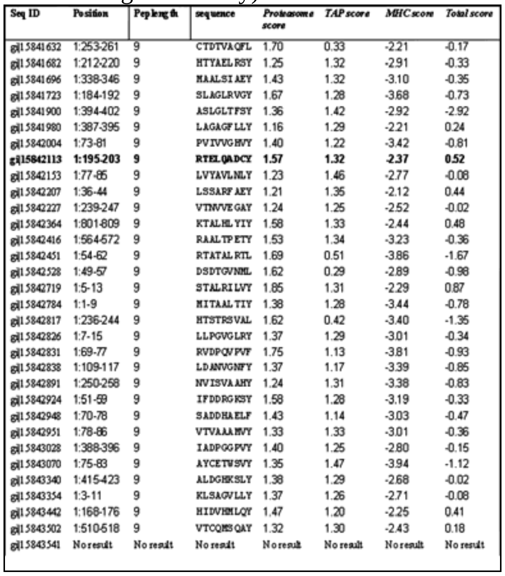
SeqID	Position	Sequence	ARB score	Smm_align score	Sturniolo score	Consensus percentile rank
gi|15841632	1:342-356	LIGPDVTREWVSDLP	1000000.0	2546.0	-1.5	57.1
gi|15841682	1:611-625	FGRTGVDCRTGPPQP	1000000.0	6664.0	-3.9	57.1
gi|15841696	1:156-170	GLGGSACTDGGKGMI	1000000.0	1935.0	-1.7	57.1
gi|15841723	1:43-57	AVAERHGRTRDEVLP	1000000.0	10229.0	-5.5	57.1
gi|15841900	1:280-294	QGQKVLHDDDNFFVA	1000000.0	47683.0	-4.4	57.1
gi|15841980	1:1-15	MSGTVVAVPPRVARA	12.4	11.0	0.9	1.6
gi|15842004	1:145-159	AADSVSDVVVDDAIT	1000000.0	29822.0	-5.8	57.1
gi|15842113	1:251-265	TTGRTNPETWTHGSA	1000000.0	6165.0	-12.5	57.1
gi|15842153	**1:7-21**	**TDRATEDHTIFDRGV**	**1000000.0**	**50000.0**	**-12.5**	**57.1**
gi|15842207	1:51-65	AILTARHDGRIVGYA	522649.4	6268.0	-5.2	53.5
gi|15842227	1:214-228	VALDDDGERHVVCSV	1000000.0	38829.0	-7.2	57.1
gi|15842364	1:748-762	LELLAERDDRITKAR	1000000.0	5513.0	-3.2	57.1
gi|15842416	1:100-114	VPADPRPKRQRITDV	1000000.0	6053.0	-6.6	57.1
gi|15842451	1:33-47	GRTVEVHVHPDDLGK	1000000.0	8734.0	-5.1	57.1
gi|15842528	1:43-57	NVFPVADSDTGVNML	1000000.0	989.0	-3.1	57.1
gi|15842719	1:108-122	EAAVPHPVDPIVLGR	98801.0	2508.0	-5.5	43.7
gi|15842784	1:67-81	LTNPSADEVKAKLVK	787264.0	815.0	-4.1	55.3
gi|15842817	1:296-310	MGPPADPECALDTLS	1000000.0	4634.0	-4.7	57.1
gi|15842826	1:67-81	GAPRIAERFTELTRE	915947.1	831.0	-6.3	55.7
gi|15842831	1:18-32	TAVTGDRDTWCVLDL	1000000.0	9996.0	-5.8	57.1
gi|15842838	1:199-213	VSMPPEADHDVAADL	843357.6	4438.0	-3.9	56.1
gi|15842891	1:73-87	RDLPDEVPVPFDVPV	1000000.0	2688.0	-3.1	57.1
gi|15842924	1:392-406	DQATKQTFTDDERAR	1000000.0	**50000.0**	-12.5	57.1
gi|15842948	1:98-112	ARFPTADCDVVRVDF	1000000.0	5624.0	-2.2	57.1
gi|15842951	1:57-71	LSEERAGLLVVRSKG	55.3	326.0	-3.5	23.9
gi|15843028	1:356-370	AGAPPGDDRVGACRQ	1000000.0	32239.0	-12.5	57.1
gi|15843070	1:34-48	IRQAPDAPDWLDAEA	1000000.0	5015.0	-3.4	57.1
gi|15843340	1:426-440	SFYTREEFDELYGGE	1000000.0	42749.0	-3.3	57.1
gi|15843354	1:109-123	WPKGSGKMRKFPEVD	1000000.0	4039.0	-2.7	57.1
gi|15843442	1:444-458	ADGWFAEAHDDSSSI	1000000.0	9666.0	-4.1	57.1
gi|15843502	1:562-576	FLVSPDGKEVIQAPY	1000000.0	2720.0	-0.8	57.1
gi|15843541	No result	No result	No result	No result	No result	No result

Table 7: MHC –I Binding Prediction Result (Low IC50 have good affinity



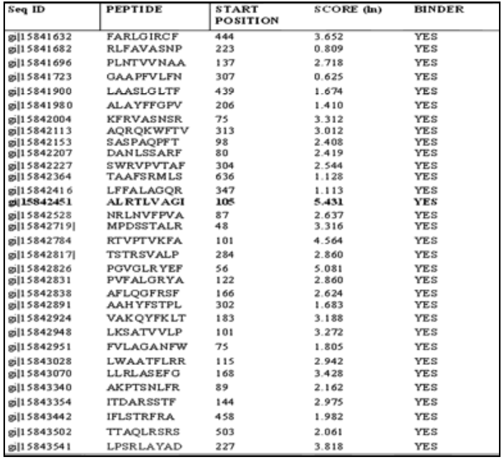
SeqID	Position	Peplength	Sequence	Ic50[n M]
gi|15841632	1:253-261	9	CTDTVAQFL	160.4
gi|15841682	1:331-339	9	LTDEAFPRL	272.1
gi|15841696	1:162-170	9	CTDGGKGMI	249.3
gi|15841723	1:239-247	9	LTALRAEMV	918.0
gi|15841900	1:207-215	9	HSDSLTAQA	477.4
gi|15841980	**1:307-315**	**9**	**LSDHSYWLV**	**74.6**
gi|15842004	1:149-157	9	VSDVVVDDA	594.5
gi|15842113	1:140-148	9	DTDFFQVLV	143.3
gi|15842153	1:64-72	9	LSDEEGLVV	125.9
gi|15842207	1:1-9	9	MTDADELAA	130.3
gi|15842227	1:276-284	9	YSDLIADWA	289.7
gi|15842364	1:410-418	9	LSAKKLARY	211.8
gi|15842416	1:111-119	9	ITDVLTLAL	280.2
gi|15842451	1:5-13	9	VVDAVEHLV	1325.1
gi|15842528	1:49-57	9	DSDTGVNML	784.8
gi|15842719	1:5-13	9	STALRILVY	194.6
gi|15842784	1:61-69	9	FADGSTLTN	865.0
gi|15842817	1:119-127	9	NADQLLIVV	1245.9
gi|15842826	1:54-62	9	LTDEEAEAV	209.2
gi|15842831	1:25-33	9	DTWCVLDLV	2281.0
gi|15842838	1:154-162	9	TTDSAPIIT	650.0
gi|15842891	1:138-146	9	LTAGVLLFT	631.1
gi|15842924	1:456-464	9	LTDTGRKYL	382.7
gi|15842948	1:1-9	9	MSAATDLYA	362.6
gi|15842951	1:57-65	9	**LSEERAGLL**	755.9
gi|15843028	1:203-211	9	HTDVLGFEA	189.6
gi|15843070	1:97-105	9	ITSPKSGVV	2658.6
gi|15843340	1:415-423	9	ALDGHKSLY	483.3
gi|15843354	1:97-105	9	ITDARSSTF	386.2
gi|15843442	1:436-444	9	LTDERIAYA	139.6
gi|15843502	1:510-518	9	VTCQMSQAY	270.9

Table 8: MHC I –Processing Prediction Result(High Score have good affinity)



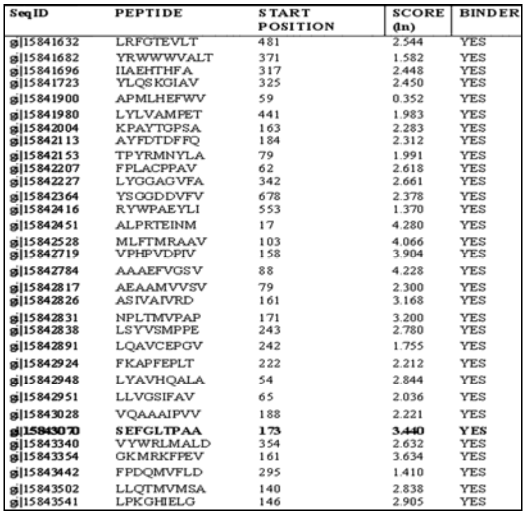
Seq ID	Position	Peplength	sequence	Proteasome score	TAP score	MHC score	Total score
gi|15841632	1:253-261	9	CTDTVAQFL	1.70	0.33	-2.21	-0.17
gi|15841682	1:212-220	9	HTYAELRSY	1.25	1.32	-2.91	-0.33
gi|15841696	1:338-346	9	MAALSIAEY	1.43	1.32	-3.10	-0.35
gi|15841723	1:184-192	9	SLAGLRVGY	1.67	1.28	-3.68	-0.73
gi|15841900	1:394-402	9	ASLGLTFSY	1.36	1.42	-2.92	-2.92
gi|15841980	1:387-395	9	LAGAGFLLY	1.16	1.29	-2.21	0.24
gi|15842004	1:73-81	9	PVIVVGHVY	1.40	1.22	-3.42	-0.81
gi|15842113	**1:195-203**	**9**	**RTELQADCY**	**1.57**	**1.32**	**-2.37**	**0.52**
gi|15842153	1:77-85	9	LVYAVLNLY	1.23	1.46	-2.77	-0.08
gi|15842207	1:36-44	9	LSSARFAEY	1.21	1.35	-2.12	0.44
gi|15842227	1:239-247	9	VTNVVEGAY	1.24	1.25	-2.52	-0.02
gi|15842364	1:801-809	9	KTALHLYIY	1.58	1.33	-2.44	0.48
gi|15842416	1:564-572	9	RAALTPETY	1.53	1.34	-3.23	-0.36
gi|15842451	1:54-62	9	RTATALRTL	1.69	0.51	-3.86	-1.67
gi|15842528	1:49-57	9	DSDTGVNML	1.62	0.29	-2.89	-0.98
gi|15842719	1:5-13	9	STALRILVY	1.85	1.31	-2.29	0.87
gi|15842784	1:1-9	9	MITAALTIY	1.38	1.28	-3.44	-0.78
gi|15842817	1:236-244	9	HTSTRSVAL	1.62	0.42	-3.40	-1.35
gi|15842826	1:7-15	9	LLPGVGLRY	1.37	1.29	-3.01	-0.34
gi|15842831	1:69-77	9	RVDPQVPVF	1.75	1.13	-3.81	-0.93
gi|15842838	1:109-117	9	LDANVGNFY	1.37	1.17	-3.39	-0.85
gi|15842891	1:250-258	9	NVISVAAHY	1.24	1.31	-3.38	-0.83
gi|15842924	1:51-59	9	IFDDRGKSY	1.58	1.28	-3.19	-0.33
gi|15842948	1:70-78	9	SADDHAELF	1.43	1.14	-3.03	-0.47
gi|15842951	1:78-86	9	VTVAAAMVY	1.33	1.33	-3.01	-0.36
gi|15843028	1:388-396	9	IADPGGPVY	1.40	1.25	-2.80	-0.15
gi|15843070	1:75-83	9	AYCETWSVY	1.35	1.47	-3.94	-1.12
gi|15843340	1:415-423	9	ALDGHKSLY	1.38	1.29	-2.68	-0.02
gi|15843354	1:3-11	9	KLSAGVLLY	1.37	1.26	-2.71	-0.08
gi|15843442	1:168-176	9	HIDVHMLQY	1.47	1.20	-2.25	0.41
gi|15843502	1:510-518	9	VTCQMSQAY	1.32	1.30	-2.43	0.18
gi|15843541	No result	No result	No result	No result	No result	No result	No result

Table 9: FDR4 Result (Affinity for HLA)



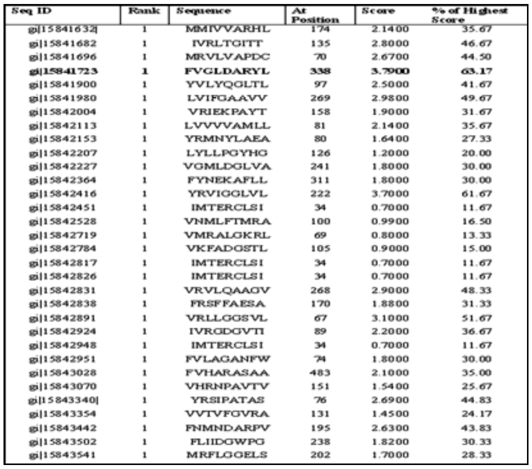
Seq ID	PEPTIDE	START POSITION	SCORE (ln)	BINDER
gi|15841632	FARLGIRCF	444	3.652	YES
gi|15841682	RLFAVASNP	223	0.809	YES
gi|15841696	PLNTVVNAA	137	2.718	YES
gi|15841723	GAAPFVLFN	307	0.625	YES
gi|15841900	LAASLGLTF	439	1.674	YES
gi|15841980	ALAYFFGPV	206	1.410	YES
gi|15842004	KFRVASNSR	75	3.312	YES
gi|15842113	AQRQKWFTV	313	3.012	YES
gi|15842153	SASPAQPFT	98	2.408	YES
gi|15842207	DANLSSARF	80	2.419	YES
gi|15842227	SWRVPVTAF	304	2.544	YES
gi|15842364	TAAFSRMLS	636	1.128	YES
gi|15842416	LFFALAGQR	347	1.113	YES
gi|15842451	**ALRTLVAGI**	**105**	**5.431**	**YES**
gi|15842528	NRLNVFPVA	87	2.637	YES
gi|15842719|	MPDSSTALR	48	3.316	YES
gi|15842784	RTVPTVKFA	101	4.564	YES
gi|15842817|	TSTRSVALP	284	2.860	YES
gi|15842826	PGVGLRYEF	56	5.081	YES
gi|15842831	PVFALGRYA	122	2.860	YES
gi|15842838	AFLQGFRSF	166	2.624	YES
gi|15842891	AAHYFSTPL	302	1.683	YES
gi|15842924	VAKQYFKLT	183	3.188	YES
gi|15842948	LKSATVVLP	101	3.272	YES
gi|15842951	FVLAGANFW	75	1.805	YES
gi|15843028	LWAATFLRR	115	2.942	YES
gi|15843070	LLRLASEFG	168	3.428	YES
gi|15843340	AKPTSNLFR	89	2.162	YES
gi|15843354	ITDARSSTF	144	2.975	YES
gi|15843442	IFLSTRFRA	458	1.982	YES
gi|15843502	TTAQLRSRS	503	2.061	YES
gi|15843541	LPSRLAYAD	227	3.818	YES

Table 10: HLA: A Binding Result



SeqID	PEPTIDE	START POSITION	SCORE (ln)	BINDER
gi|15841632	LRFGTEVLT	481	2.544	YES
gi|15841682	YRWWWVALT	371	1.582	YES
gi|15841696	IIAEHTHFA	317	2.448	YES
gi|15841723	YLQSKGIAV	325	2.450	YES
gi|15841900	APMLHEFWV	59	0.352	YES
gi|15841980	LYLVAMPET	441	1.983	YES
gi|15842004	KPAYTGPSA	163	2.283	YES
gi|15842113	AYFDTDFFQ	184	2.312	YES
gi|15842153	TPYRMNYLA	79	1.991	YES
gi|15842207	FPLACPPAV	62	2.618	YES
gi|15842227	LYGGAGVFA	342	2.661	YES
gi|15842364	YSGGDDVFV	678	2.378	YES
gi|15842416	RYWPAEYLI	553	1.370	YES
gi|15842451	ALPRTEINM	17	4.280	YES
gi|15842528	MLFTMRAAV	103	4.066	YES
gi|15842719	VPHPVDPIV	158	3.904	YES
gi|15842784	AAAEFVGSV	88	4.228	YES
gi|15842817	AEAAMVVSV	79	2.300	YES
gi|15842826	ASIVAIVRD	161	3.168	YES
gi|15842831	NPLTMVPAP	171	3.200	YES
gi|15842838	LSYVSMPPE	243	2.780	YES
gi|15842891	LQAVCEPGV	242	1.755	YES
gi|15842924	FKAPFEPLT	222	2.212	YES
gi|15842948	LYAVHQALA	54	2.844	YES
gi|15842951	LLVGSIFAV	65	2.036	YES
gi|15843028	VQAAAIPVV	188	2.221	YES
gi|15843070	**SEFGLTPAA**	**173**	**3.440**	**YES**
gi|15843340	VYWRLMALD	354	2.632	YES
gi|15843354	GKMRKFPEV	161	3.634	YES
gi|15843442	FPDQMVFLD	295	1.410	YES
gi|15843502	LLQTMVMSA	140	2.838	YES
gi|15843541	LPKGHIELG	146	2.905	YES

Table 11: ProPred MHC Binding Peptide Prediction result



Seq ID	Rank	Sequence	At Position	Score	% of Highest Score
gi|15841632|	1	MMIVVARHL	174	2.1400	35.67
gi|15841682	1	IVRLTGITT	135	2.8000	46.67
gi|15841696	1	MRVLVAPDC	70	2.6700	44.50
gi|15841723	**1**	**FVGLDARYL**	**338**	**3.7900**	**63.17**
gi|15841900	1	YVLYQGLTL	97	2.5000	41.67
gi|15841980	1	LVIFGAAVV	269	2.9800	49.67
gi|15842004	1	VRIEKPAYT	158	1.9000	31.67
gi|15842113	1	LVVVVAMLL	81	2.1400	35.67
gi|15842153	1	YRMNYLAEA	80	1.6400	27.33
gi|15842207	1	LYLLPGYHG	126	1.2000	20.00
gi|15842227	1	VGMLDGLVA	241	1.8000	30.00
gi|15842364	1	FYNEKAFLL	311	1.8000	30.00
gi|15842416	1	YRVIGGLVL	222	3.7000	61.67
gi|15842451	1	IMTERCLSI	34	0.7000	11.67
gi|15842528	1	VNMLFTMRA	100	0.9900	16.50
gi|15842719	1	VMRALGKRL	69	0.8000	13.33
gi|15842784	1	VKFADGSTL	105	0.9000	15.00
gi|15842817	1	IMTERCLSI	34	0.7000	11.67
gi|15842826	1	IMTERCLSI	34	0.7000	11.67
gi|15842831	1	VRVLQAAGV	268	2.9000	48.33
gi|15842838	1	FRSFFAESA	170	1.8800	31.33
gi|15842891	1	VRLLGGSVL	67	3.1000	51.67
gi|15842924	1	IVRGDGVTI	89	2.2000	36.67
gi|15842948	1	IMTERCLSI	34	0.7000	11.67
gi|15842951	1	FVLAGANFW	74	1.8000	30.00
gi|15843028	1	FVHARASAA	483	2.1000	35.00
gi|15843070	1	VHRNPAVTV	151	1.5400	25.67
gi|15843340|	1	YRSIPATAS	76	2.6900	44.83
gi|15843354	1	VVTVFGVRA	131	1.4500	24.17
gi|15843442	1	FNMNDARPV	195	2.6300	43.83
gi|15843502	1	FLIIDGWPG	238	1.8200	30.33
gi|15843541	1	MRFLGGELS	202	1.7000	28.33

MODELLED STRUCTURE

The 3D structure of the sequence with a score value more than 35% were modeled. List of modeled protein sequences and their templates are listed in the table 12

List of Modeled hypothetical proteins having predicted functions.

ID	Percentage of identity	Pdb id
NP_337985.1	43%	2BVF
NP_337961.1	37%	3FCR
NP_337453	98%	2HYX
NP_336733	35%	3CWC

Modeled 3-D structures of hypothetical sequences

3-D structures of the modeled proteins of hypothetical sequences are shown in figure-1. These modeled structures have been used for vaccine region prediction.

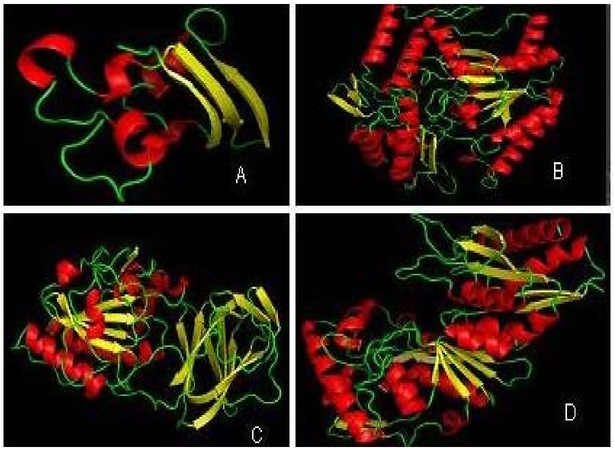


Fig 1: Modelled structure of hypothetical protein (A) NP_337985.1 having function FAD/FMN-containing dehydrogenas (B) NP_337961.1 having function Adenosylmethionine-8-amino-7-oxononanoate aminotransferase. (C) NP_337453.1 having function Cytochrome c biogenesis protein. (D) NP_336733.13 having function Glycerate kinase

SWISS PDB VIEWER VISUALIZATION

Swiss PDB viewer was used to find out the antigenic and other vaccine coding region MHC-I, MHC-II, HLA, TAP region and the marked antigenic regions are shown in figures-2

.

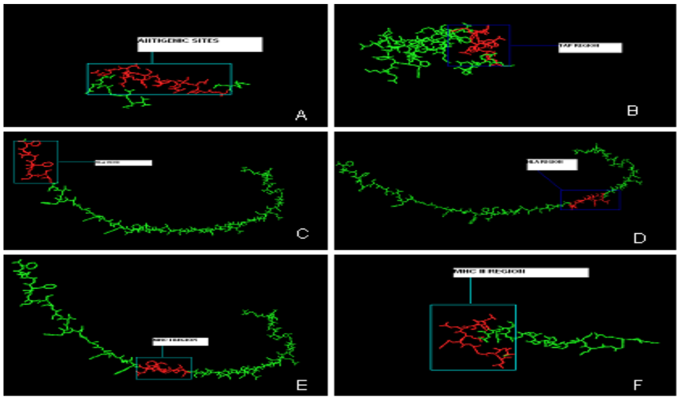


Figure-2 Antigenic region for (A) NP_337988.1, (B) TAP regions for NP_337401 (C) HLA regions for NP_337985.1(D) HLA region for NP_337985.1 (E) MHC I region for NP_337985.1.(F) MHC II region for NP_337988.

DISCUSSION

Tuberculosis is a contagious, deadly infectious disease caused mainly by *M. tuberculosis*. Tuberculosis usually attacks the lungs (as pulmonary TB) but can also affect the central nervous system, the lymphatic system, the circulatory system, the genitourinary system, the gastrointestinal system, bones, joints, and even the skin. The main causative organism for tuberculosis disease is *M. tuberculosis* CDC1551 species. So, understanding the genome of *M. tuberculosis* will help us to take necessary steps to avoid or prevent tuberculosis disease. While observing the genome of *M. tuberculosis* as functional and non-functional categories, it eas found that almost 50% percent of the sequence does not possess any function. Manual re-annotation of the whole genome of *M. tuberculosis* helped to produce 37 % new coding sequences with functions.

From the annotation results it has been observed that out of 1,070 hypothetical sequences or ORF's only 32 hypothetical sequences share 100% functional identity. The presence of secondary structure like helix and sheets and the tertiary structure in the modeled proteins could contribute a significant role in the lipid metabolism of *M. tuberculosis*.

Further, Reverse vaccinology work helps to predict the vaccine regions for those newly identified protein sequences. Of all the 32 hypothetical protein sequences taken for vaccine studies only the sequence with I.D NP_337985.1 shares structural epitope region with all the four epitope regions including TAP region, MHC-I and II binding region, HLA region. Other sequences with I.D gi_15843442, gi|15842451, gi|15842113, gi|15841980 shares the sequential epitope region.

The average length of epitope region is found with the help of sequence analysis and is ~15 amino acids. Hence, these small stretched sequence patterns of *M. tuberculosis* may have the antigenic role in human immune system.

CONCLUSION

Reverse Vaccinology stands as a turning stone in Vaccinology. Reverse vaccinology prediction work can be used on a large number of bacterial and viral proteomes are reliably effective in selecting probable vaccine candidate pools that can be characterized as an antigen.

From this work, it can be found that the hypothetical sequence NP_337985.1 has a function as FAD/FMN-constaining dehydrogenase which shares the entire epitope region with high scoring value. So, it is clearly inferred that this protein can be act as a better vaccine that can act against *M. tuberculosis*. This work will aid researchers in designing subunit vaccines that might cure tuberculosis disease.

ACKNOWLEDGEMENT

One of the author P.Vijayalakshmi kindly cknowledge DBT for research fellowship and for providing bioinformatics infrastructure facility.

REFERENCES

[1] Aboa-Zeid C., Smith I., Grange JM., Ratliff TL., Steele J., Rook GAW(1998) The secreted antigens of *Mycobacterium tuberculosis* and their relationship to those recognized by the available antibodies: J Gen Microbiol 134-531.

[2] Silva C. (1995) New vaccines against tuberculosis: Brazilian Journal of Medical and Biological Research 28:843-851.

[3] Lankat-Buttgereit B., R.Tampe (1999) The transporter associated with antigen processing TAP: structure and function. FEBS Lett. 464 :108-12.

[4] Lundegaard C., Lamberth K., Harndahl M., Buus S., Lund O., and M.Nielsen(2008) NetMHC-3.0: Accurate web accessible predictions of Human, Mouse, and Monkey MHC class I affinities for peptides of length NAR 36:509-512.

[5] Tenzer S., Peters B., Bulik S., Schoor O., Lemmel C., Schatz M.M., Kloetzel P.M., Rammensee HG., Schild H., and Holzhutter H.G.,(2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding: Cell Mol Life Sci. 62:1025-1037.

[6] Parker K. C., Bednarek M.A., and J. E. Coligan(1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains: J. Immunol. 152:163.

[7] Singh.H. and G.P.S.Raghava (2001) ProPred: Prediction of HLA-DR binding sites: Bioinformatics,17:1236-37.

[8] Eswar.N., Marti-RenomM.A., Webb B., Madhusudhan M.S., Eramian.D, Shen.M., Pieper.U., A. Sali (2000) Comparative Protein Structure Modeling With MODELLER. Current Protocols in Bioinformatics: 15:5.6.1-5.6.30